Journal of Autonomous Intelligence (2023) Volume 6 Issue 3
doi: 10.32629/jai.v6i3.726

REVIEW ARTICLE

Efficient integration of big data with blockchain: Challenges, opportunity and future

Himani Saraswat[1], Sanjay Jasola[2], Mahesh Manchanda[1,*]

[1] *Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun 248001, India*

[2] *Department of Computer Science and Engineering, Graphic Era (deemed to be) University, Dehradun 248001, India*

*** Corresponding author:** Mahesh Manchanda, maheshmanchanda@gmail.com

ABSTRACT

Big data has become more and more popular, piquing the interest of both researchers and technologists as well as business executives. Despite its advantages, big data has a number of problems that necessitate a one-stop shop. Blockchain technology has seen a considerable increase in usage, which has had a substantial impact on its applications and led to a variety of useful outcomes. In the areas of identity, trust, decentralization, data-driven decisions, data ownerships, etc., there are notable game-changers. As a result, Blockchain is frequently acknowledged as an effective fix for big data issues. Among the solutions it suggests are decentralized private data management and digital property resolution. Together, two these technologies can develop beneficial solutions.

Keywords: blockchain; big data; projects; decentralized; security; data

ARTICLE INFO

Received: 15 June 2023
Accepted: 19 July 2023
Available online: 18 September 2023

1. Introduction

As the world expands its reach on a variety of fronts, so does the amount of data created and authorized. The digitalization of the world is responsible for such high levels of data generation. The traditional data always had a block consisting of documents, texts, finances, stock records, and personal files, but now the data is becoming more complex as it travels faster. The main concern is how this fast-moving data can be secured, manipulated, and delivered to its destination in an intact state. The amount of data transmitted over the internet expands at an exponential rate every year. Big data is defined, according to a McKinsey Global Institute report, a new generation of technologies and architectures are being researched to assess vast amounts of data and capture their key properties[1]. Retailers can enhance prospective operating margins by 60%. (e.g., high velocity, knowledge discovery, and analytics). Big data is defined as datasets of extraordinary size and dimensionality that can't be stored, handled, analyzed, or collected with typical database methods.

Every year, transmitting data increases at a considerably quicker rate than the typical variation. The demand for more complicated data is growing by the day. Security and fault tolerance are additional focuses of the complexity. The amount of data being streamed in real time is enormous, and preserving, storing, and protecting it is a difficult undertaking. To have brought up to the duty of this evolution are the projects or companies, those who work in these fields. The

market trends and forecasts shown by this data can help accelerate the planning process and reveal patterns that can be used to action to develop.

The procedure of the presentation, management, restoration and the security of this data has always been a concern to the researchers. The key objective is to figure out how to manage the computation of this data while also managing it. Data security falls under this broad category and is limited to issues such as data confidentiality, availability, and integrity maintenance. The term "data confidentiality" relates to the possibility of illegal access to this information, whereas "data availability" refers to the availability of this information to the right and authorized users. This paper contributes towards the prospectives of big data and BCT for the future researches'. The research creates possibilities of fixing maximum issues of big data from BCT in terms of security, management, and quintessential future scope of researches. This paper is divided in five sections, the first section is about big data, then second section is introduction to blockchain the third section is about big data and blockchain the next section is about the big data projects and how blockchain is creating its impact in the challenges in them. Followed by the discussion, future scope and at the end conclusion.

The gathering, analysis and application of big data analytics, which has permeated every aspect of daily life and is advancing quickly, have been greatly aided by analytics[2]. Big data is defined as information of extraordinary size and dimensionality that can't be stored, handled, analyzed or collected with typical database methods. It was presented a survey of large data management strategies and technologies[3]. The volume, velocity, and noticeable representation, as well as the requisite scaling methods, all lead to big data having an architectural point of view. In terms of data security, privacy, energy management, data management, interpretation and real-time data, big data presents a lot of challenges and issues. Data security, privacy, energy management, data management, interpretation, and real-time data processing are all issues that need to be addressed. Two of the most realistic possibilities are privacy and security. Various data secrecy preservation systems and techniques have been considered, looked into using reinforcement learning to generate a security-aware procedure for a smart grid system[4]. **Table 1** shows the comparison between standard data with big data to help better grasp what big data is.

Table 1. Comparison in traditional and big data.

Traditional data	**Big data**
Credentials	Pictures
Assets	Auditory & audio visual
Stock records	3D model
Personnel records	Replications, position data

Companies can take their development to the next level by gathering data in a digital format. Digital data analysis can aid in the planning process by revealing trends that can be used to improve strategy. Receiving real-time consumer demand information is useful for observing market patterns and predicting. The term "big data" also refers to how information is managed. There must be a collection of tools that can go through and filter vast amounts of data that is exceedingly complex and varied while processing massive amounts of data. Predictive analytics can be performed with big data, which is something that many firms rely on to figure out how they're heading[5]. For example, a telecoms company can use information such as call length, average text messages sent, and average bill amount to estimate which customers are more likely to discontinue their services.

Big data is typically characterized by the 4 V's, which make it more reliable to consider; velocity, volume, variety and veracity. It might be a massive assortment of data that is growing suggestively over time. It is a dataset that is so enormous and intricate that no typical data management technologies can effectually

store or process it[6]. Big data analytics plays an important role in the implementation and creation of a turf field for analytical field. BDA (Big Data Analytics) is about mining useful evidence and considering outlines from the stated dataset, which can be used for dissimilar purposes for business and research. The best example of the big data is the data generation in stock exchange but also to many verticals like e-health, smart grid, mobile data logs, transportation and logistics. The term "volume" refers to the amount of data being modified and analyzed in order to achieve the intended outcomes. It is a difficult task since in order to handle and evaluate a large amount of data necessitates a large amount of resources, which will cost a lot of money finally result in the appearance of the outcomes that were requested. The amount of data that can be processed is limitless, but the speed at which it is processed is constant. To attain better processing speeds, more computer power is required, which necessitates the development of infrastructure, but at a larger expense. Velocity, due to the growing demand for streaming data across several devices by end users, this is also a major concern. This is a difficulty that the majority of businesses are unable to meet. Characteristically, data is transferred at a rate that is less than the system's capacity. Because transfer speeds are finite but demands are limitless, streaming data in real-time or near-real-time is a significant difficulty. Variety the representation of it denotes the type of information that is saved, processed, and utilized. Location coordinates, video files, data provided from browsers, simulations, and other types of data can be kept and evaluated. The problem is figuring out how to arrange all of this information so that it is "readable" by all people who access it and does not produce ambiguous outcomes. The value and veracity refer to quality of data generated at the real time. If there is a loss of data from one geo-location, it is not an issue in big data because there are hundreds more that can cover that information.

As a ledger technology, blockchain has emerged as one of the most striking solutions for implementing security and privacy in big data systems. According to researchers blockchain has played a critical role in providing high-quality data and securing data sharing for industrial IoT applications. Researchers proposed a blockchain-based machinery for securing mobile data collection and incentivizing mobile nodes for efficient data collection. In many cases it is found that blockchain is found in edge computing to participate the data quality and process the computation initiated tasks.

2. Blockchain: The overview

The father of Blockchain Satoshi Nakamoto[7] introduced the cryptocurrency bitcoin and the basic blockchain technology (BCT) have created a significant buzz around electronic payment solutions that use the internet's peer-to-peer paradigm[8]. The blockchain is a shared, distributed, decentralized, immutable and secure data structure. Blockchain can be used as a database, but also as a platform which determine protocol for establishing consensus without a central hub or any intermediary institution. When a transaction is carried out between two nodes on the bitcoin network, it is broadcast throughout the whole network[9]. If the transaction is genuine, every node in the network verifies it and adds it to their transaction pool; if not, they reject it. Each node keeps a separate transaction pool. The transaction pool contains all of the legitimate transactions.

In simple terms, blockchain technology is a distributed ledger technology; it is a shared public/private ledger of all digitally extended[10] events that have been executed and shared among blockchain stakeholders. A stakeholder in blockchain technology creates a new transaction to be added to the blockchain. A single record of data is stored in a block on each stakeholder's node in blockchain. The list of information records (blocks) is highly encrypted, ensuring a high level of privacy and security. This technology combines cybersecurity[11], cryptography, software engineering[12], and distributed computing[13] It is not a hyperbolic statement to say that the present and future of the industry are based on financial transactions between stakeholders; it promotes transparency and security[14].

A blockchain is a distributed database that stores encrypted blocks of data that are then linked together to form a single-connection-trust of data[15]. Blockchain is a cutting-edge and revolutionary technology that reduces risk, eliminates fraud, and increases system transparency. Financial theft has become much more difficult as a result of blockchain technology[16]. Blockchain is upending the current state of innovation by enabling businesses to experiment with cutting-edge technology like peer-to-peer energy distribution and decentralized news delivery. When a sender initiates a Bitcoin transaction, it is sent to the receiver via a transaction completed on the public bitcoin network. Users are verified by network miners, who also confirm that the sender has enough bitcoins to send to the recipient[17]. There are three types of blockchain: public, private and hybrid. The public blockchain is where Bitcoin and other cryptocurrencies like it were born, and where they contributed to the advancement of distributed ledger technology (DLT). It eliminates centralization's drawbacks, such as a lack of security and transparency[18]. Anyone with internet access can join a blockchain platform and become an authorised node, making the public blockchain open and permissionless. A blockchain network is made of the following layers:

- Hardware layer: The server that hosts the entire network or the cloud might function as the hardware layer.
- Ledger or fabric layer: This is the building block of the blockchain network and is made up of blocks. Since it contains transaction information, it is sometimes referred to as a ledger.
- Smart contract or logic layer: This establishes the network's business logic and guarantees that it complies with all laws and guidelines governing the blockchain network.
- Interface layer: This group of APIs (Application Programming Interface) is used to interact with the blockchain and obtain the desired outcomes, such as data addition and retrieval.
- User-interface or application layer: This layer controls the entire network and serves as the application's front end. It engages in interaction with the other layers.

Private blockchain; a private blockchain is a blockchain network that operates in a restricted context, such as a closed network, or is controlled by a single entity. While it functions similarly to a public blockchain network in terms of peer-to-peer connectivity and decentralization, this blockchain is substantially smaller. Hybrid blockchain; a type of blockchain technology that incorporates features from both private and public blockchains[19]. It enables businesses to create a private, permission-based system alongside a public, permissionless system, allowing them to control who has access to specific data stored on the blockchain and what data is made public. A blockchain that is run and maintained by a single entity is considered private. These kinds of blockchains are typically appropriate to conglomerates, where the parent firm controls the network for the underlying collection of businesses. They priorities efficiency over immutability, anonymity, and openness in certain circumstances. The RBI (Reserve Bank of India) might be seen as the organization with ultimate control over the entire network if we take the India Lending Blockchain into account. But this poses the issue of giving one organization an excessive amount of control.

While a consortium blockchain and a private blockchain have many similarities, they differ when it comes to network management and control. Authority is divided between two or more entities rather than being centralized on one. This scenario is also suitable for the India Lending Blockchain where authority can be distributed between RBI (Reserve Bank of India) and a few of the major banks so that benefits can be ensured for all the members.

Immutability, decentralization, security, distributed ledger, consensus, and faster settlement are all characteristics of blockchain technology.

a) Immutability: the technology without any corruption; a permanent, unaltered network. In terms of any sort of transaction happening the nodes check the authenticity and add it in ledger[19]. So, it can be said the success of any sort of transaction can be only possible. The transactions are visible to everyone on the public

blockchain, making it extremely transparent. Private or federated blockchain, on the other hand, may be best for businesses that want to maintain transparency among employees while also protecting sensitive information from public scrutiny[20,21].

b) Decentralized: It means that it has no governing authority and is maintained by a group of nodes, making it decentralized[22]. Because the system does not require any governing authority, stakeholders can access it and store data directly from the web.

c) Security: Every piece of information on the blockchain has been cryptographically hashed. Simply put, network information hides the underlying nature of the data[23]. Any input data is passed through a mathematical method that generates a different type of value while keeping the length constant. It can be thought of as a one-of-a-kind data identification.

d) Distributed ledger: All other system users contribute to the network ledger. To achieve a better result, computational power was distributed across the computers. This is why it is regarded as one of the most important characteristics of the blockchain. The end result is always a more efficient ledger system capable of competing with the old ones. Every active node is required to maintain the ledger and participate in validation, as well as the intermediates[24].

Mining refers to the process of creating new blocks and adding them to the blockchain. Miners create blocks from the transactions present in their transaction pools. When a miner approves and verifies a transaction, it is added to the block and becomes part of the blockchain network. Finally, the block's relevant transactions are carried out, updating ledgers across all nodes and ensuring that all participants have the same copy of the transaction for transparency and security. The ledger system's applications, like the concept of blockchain, will only expand as technology advances. Blockchain technology can be used to track financial crime, securely transmit patient medical records among healthcare professionals, and even track intellectual property in the corporate world and music rights for musicians. A blockchain is a reliable and tamper-evident system since it is immutable. This becomes a constraint, too, because it might be necessary to amend some inaccurate data in a transaction. Organizations have developed a method known as chameleon hash to get around this restriction.

It enables a participant to change the base transaction data without changing the block's hash. The hash value of the amended block and the subsequent blockchain would be unchanged, but the modified block would have a scar to show that the contents was changed. Future developments are anticipated to solve similar issues with more of these evolutions. **Table 2** summarizes, how all type of blockchain behave in different parameters.

Blockchain employs a set of validation checks, known as a consensus mechanism, to check the validity of transactions. A consensus algorithm is a computer science procedure that allows disparate processes or systems to agree on a single data value. Consensus methods are used in networks with several faulty nodes to achieve reliability. In distributed computing and multi-agent systems, resolving this issue, known as the consensus problem, is critical. The main difference between the two blockchain type is basically the potentially applied consensus algorithm. The different type of consensus algorithm are Proof-of-Work (PoW), Proof-of Stake (PoS), Practical Byzantine Fault Tolerance (PBFT), Delegated Proof of Stake (DPoS). The PoW has made a tremendous success in bitcoin, PoW requires contributors that compete for mining blocks to give the proof of their work. The general mathematical expression for the proof of work algorithm is:

$$\text{SHA256 (SHA256 (h. n))} \leq \text{target}$$

A block header hash is calculated as a double SHA256 hash of all the block constituents, as shown below:

Block header hash = SHA256 (SHA256 (Previous block hash + Merkle root + Timestamp + Difficulty target + Nonce))

where h is the content of the newest block and it is always seen smaller the target; the difficulty level of mining is high.

The next Proof-of Stake (PoS) in this the main idea is to compete mining. It has been seen as a replacer for PoW, as it requires less of the computational resources[25]. The Hash is like a fingerprint of the block; a 64 character long cryptographically linked with each other It is basically constituted to identify some document or content of the newest block. Nounce, is the valid number of the hash block. The network's difficulty is determined by the difficulty value, also known as the target value. The network's difficulty level is used to control block mining. The bitcoin network has a 10 min block creation time that must be maintained. With each block construction, the difficulty value adjusts so that the block creation value remains constant.

Table 2. Summarizes the type of blockchain and their different parameter to commiserate.

TYPE	Anonymity	Transparency	Immutability	Efficiency	Confidentiality	Throughput	FAT
Public	Y	Y	Y	N	Low	Low	High
Permissioned	N	N	Y	Y	Medium	Medium	Medium
Private	N	N	N	Y	V. High	High	Low
Consortium	N	Partial	Y	Y	High	High	Low

3. Big data meets blockchain

Both government and private organizations are investing in big data remediation. Management, data cleansing, imbalanced system capacities, imbalanced data, analytics, and learning from data are the challenges and issues in big data. The use of blockchain has the potential to solve a wide range of real-world problems. Throughout today's world, customers prefer to transact online, and the growing amount of data being generated opens new opportunities for industries to better understand customer needs, purchasing patterns, and trends. The rate of most complex problems is increasing day by day. Streaming real-time information is a challenge that businesses must overcome. Cryptocurrencies operate in a system that cannot be manifested as an encrypted digital currency, and the massive overall network's well-structured comprehensive records satisfy the 5 V feature of big data (volume, variety, velocity, veracity, and value).

Because the ledger is open to the public, anyone can view the blocks and transactions. The users, on the other hand, maintain their anonymity by identifying themselves solely using their public key as an address. Transactions are encrypted as well. Transactions that are invalid are rejected and do not appear in blocks[18]. Malicious modifications to the transactions will necessitate recalculating the proof of work for the attached block and all subsequent blocks. These computations are impossible unless most nodes in the network are malicious.

Users can use blockchain transactions to store or directly exchange their valuables. In the case of bitcoin, the asset is a digital currency, but blockchain transactions are not limited to that use. They can represent physical 0 or digital property, a smart contract between two or more parties, or any other type of data or document[26]. The combination of blockchain and smart contracts has the potential to create a new generation of transactional applications that prioritize trust, accountability and transparency while streamlining corporate operations and avoiding legal constraints.

The motivation for having a integration of blockchain along with big data is:

a) Data security and privacy: The web is causing a revolution in today's world, and the amount and quality of data stored in third-party locations such as cloud storage is go through the ceiling[23]. Conventional protection mechanisms, such as firewalls, have evolved over time, as data storage locations have expanded beyond the organization's perimeters. One possible solution is to use the blockchain. The decentralized storage and encryption make it more difficult for any unapproved access to the data.

b) Data reliability: People may tamper with big data records in order to sway big data analytics predictions in their balance in favor[24]. The immutability of the blockchain assures that meddling with data stored in the blockchain network is practically impossible. In the event of data alteration, they must change at least half of the data.

c) Fraud detection: To detect fraudulent transactions, current big data systems rely on pattern analysis of previous data. Financial organizations can track every financial transaction in real time because data is kept in blockchain[25]. As a result, the blockchain can prevent financial institutions from being fraudulently preventive.

d) Relevant data analytics: Real-time data analytics is more practical because the blockchain records every contract. Because the blockchain integrated big data analytics allows the financial institution to settle transactions fast, it has 'cross-border' transactions in real-time.

Figure 1 depicts an outline of blockchain services in the big data environment in terms of big data processing, storage, analytics, and big data privacy-preserving. The data acquisition block uses data from diversified sources in a unstructured format. The IOT devices exchange the data transmission in terms of sharing and prevention of data theft. The data storage (example, HDFS, IPFS, EduRSS and BigData-as-Service) and access files from anywhere on any machine. Blockchain is used to record the data storage and retrieval process. The hash value is stored in the blockchain for authenticity of user verification. These things are depicted by the **Figure 1**. Data can be evidenced from a variety of sources, including data reports, data libraries, social media, and assistive devices. The Blockchain mining section contributes to block creation, data mining, consensus algorithms and smart contract written. The vast expansion of the cyber-physical system enables speedier information services, as well as real-time sensing and access. Data from big data is sent to a cyber-physical system that communicates over radio waves. Spectrum auctions are fiercely competitive, and license-free spectrum is scarce.

When data is huge the blockchain has to deal with the huge amount of this real time data along with the high amount of throughput. Anyone dealing with data should be aware of the big data analytics. The blockchain have the ability to search-and-retrieve policy. **Table 3** provides an overview of the services provided by big data, its obstacles, and how blockchain technology has provided solutions to the most egregious big data problems; the paper citations and author are also mentioned.

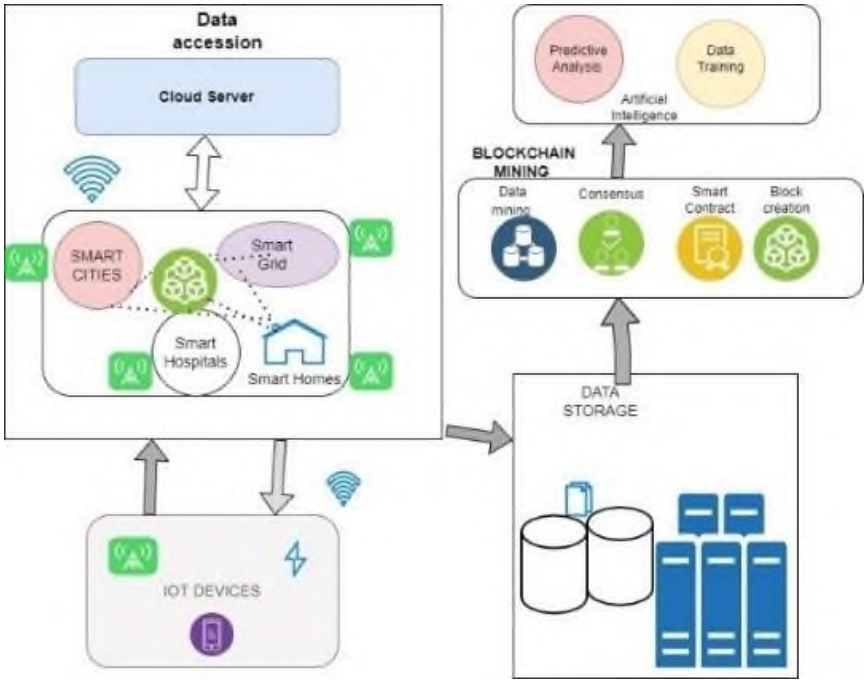

Figure 1. Blockchain in big data environment.

Table 3. Services and solutions.

Services for big data	Challenges	Solution by blockchain	Reference
Data collection	Data subjected to malicious attacks, unprecedented methods and threats	Supports efficient energy data, data allocation like; Ethereum, Hyperledger fabric	[27]
Data share	Lack of authorization, response time	The unproductive transaction algorithm accesses data from the cache layer in order to reduce response time and processing head. Smart contracts are used to grant permissions.	[28]
File system (storage facility)	Unauthorized access, privacy, redundancy security, duplicity.	Coupled with IPFS, which gives a solution by employing decentralized platforming to solve file redundancy difficulties while also providing file storage security. Hashing values are stored in blockchain to provide the user with authenticity, and the encryption method is used before cloud storage.	[29]
Database management system	Data preserved in a distributed database is vulnerable to both internal and external threats.	Data tampering is not possible in blockchain, use of time stamp overcome this issue. The virtual sharing ledger is applied to store the transaction history data. The transaction done in database are stored in inter-connected blocks using cryptography.	[30]
Data training	Many entities are to share data for integration in training and splitting datasets for many ML classification algorithm.	Cryptosystems provide a secured environment that does not require the intervention of a trusted third party.	[31]
Privacy	Data privacy issues occur during sharing of data from different sources and entities. User privacy is also at stake. Many third party stakeholders are exposed in ways of security breaches and in regards to data misuse.	Gives immutable, verified, and decentralized ledger to record the transaction in digital scenarios. Crypto-privacy is applied for solving privacy preservation.	[32]

The combination of big data and blockchain can open up a number of intriguing possibilities for measured modelling. Here is a scientific modelling strategy for analysing large amounts of data utilising blockchain:

- Data representation: Decide the manner in which you want to portray your large amounts of data. This may entail establishing the data structure, identifying important variables, and selecting acceptable data types. This stage is critical for successfully organising and processing data.
- Data preprocessing: To deal with missing values, outliers, and noise in big data, preprocessing is frequently required. Consider using data cleaning, normalisation, dimensionality reduction, and feature selection techniques to guarantee high-quality data for analysis. Blockchain integration: Determine which features of blockchain technology you want to include in your computational framework. This might include utilizing the blockchain's immutability and transparency capabilities, as well as smart contracts for automated data validation and execution.
- Model development: Choose a theoretical framework or method that is compatible with your goals and data characteristics. Regression models, classification models, clustering methods and time series analysis approaches are all popular options. If required, modify the model to fit the blockchain environment.
- Model training and validation: Separate your data into training and testing sets. Train the model using training data and assess its performance with testing data. For more robust validation, consider approaches such as cross-validation or bootstrapping.

- Integrate blockchain verification techniques: Incorporate blockchain verification techniques into your model. This may include the use of cryptographic techniques to secure data integrity and consensus procedures to obtain agreement on the data's authenticity.
- Performance evaluation: Evaluate your mathematical model's performance in terms of accuracy, precision, recall, F1 score, or other applicable evaluation metrics. To assess the model's efficacy, compare the outcomes to benchmarks or prior models.
- Iterative refinement: As needed, iterate and refine your model. Analyse the findings, identify areas for improvement and modify your strategy appropriately. This may entail fine-tuning model parameters, integrating more variables, or experimenting with different methods.

The performance measure for blockchain indulge in ensuring the big data training and prevent data theft to facilitate big data transmission. Data can be recorded from pervasive sources like reports, social media libraries. They are added to blockchain with digital signature and hash value before sharing with data analytics services in which both data source owner and data analytics user can trace and monitor the data sharing flow.

4. Blockchain—Big data in projects

Blockchain technology has earned significant praise for its numerous applications in a variety of real-world problems. Because technology is still in its infancy, many of the difficulties to be addressed, such as data ownership and decision support systems, are still in their infancy. Our modern world is awash in massive amounts of data that is generated or created by both humans and machines[33]. The increased demand for storage, organization, processing, and analysis necessitates a significant role for blockchain[34]. Several machine learning (ML)[35] and deep learning (DL) methods are employed for successful data analysis. Because of its efficiency and precision, the support vector machine (SVM) is a popular machine learning (ML) approach[36]. In the framework of vehicle-to-vehicle social networks data is gathered from several sources, including social media sites. Automobile manufacturers, companies, and vehicle management services are all involved[37]. **Table 4** summarizes the services and initiatives available. Blockchain technology is utilized in data analytics applications to examine trade trends, anticipate new clients, diseases and business partners.

Table 4. Summarizes the project and challenges.

BDA on projects	Technique/description	Advantages	Research challenges	Reference
Smart cities	The use of a hash. Encryption that is asymmetric. Merkel tree and consensus algorithm. Big data auditing. Fog nodes & D2D blockchain.	Alteration free transaction in IOT device. Decentralized technique brings secure IOT device environment. Risk prediction is easy	Maintenance of balance in privacy. Technique to find crowd sensing.	[38]
Smart health care	Patient record accessed by cryptographic technique. Implementation of IoT based EHR system. Ethereum smart contract used for communication of sensors with smart devices.	Provide security and privacy for the health monitoring system inspired by IoT. Safeguard from fraud detection and identity verification. Monitoring of patient data in real time and sending of alerts for medical intercession.	Commercialization in regards industrial partner. Large-scale healthcare data implementation. Difficulty in utilizing resources.	[39]
Power grids	Consensus algorithm and hashing implementation of crypting algorithm in data integration and regulatory system. Use of blockchain secured response management system.	P2P energy trading and a decentralized energy generation system. Smart grid security is achieved through the use of a multi-key scheme.	Difficulty in scalability in process transaction. Price prediction is difficult. Realtime analysis is little difficult to be included.	[40,41]

The use of blockchain in large data initiatives can improve data integrity and prevent sensitive data from being tapered. Though many applications confront research obstacles, evaluating trading trends, projecting new clients, diseases, and business partners are all advantageous. Many blockchain-enabled big data analysis initiatives are gaining popularity in the market, like Omnilytics and Rublix, and datum is well-known for anonymously storing data.

5. Discussion and future prospects

Emerging blockchain in big data platform services and applications is emphasized and analyzed in this study. The broad literature research uncovered certain significant technological difficulties that must be addressed, as well as techniques. The integration of blockchain into bigdata applications, services and projects is fraught with difficulties. The answer to issues such as data security, the blockchain technique interprets data exchange only to authorized entities, which is a corrective option for decentralized data management and data exchange concerns[42]. Unauthorized access is a problem in big data file management, but combining blockchain with a decentralized platform to handle file redundancy provides security to the file storage system. It has concerns with external and internal threats as a database management system for huge data, but blockchain resists data tampering and uses data stamping methods. In a digital situation, blockchain provides an immutable, decentralized ledger for transaction records. There are some open research challenges in the integration of BCT and big data[43]:

A) Privacy: Many private blockchain platforms will limit access, exposer and credentiality of the large data set, which can be necessary for BD (Big Data) to process along with AI and reach to correct decision[44].

B) Fog computing paradigm: The fog nodes must be equipped with AI and machine learning capabilities, as well as a blockchain interface, allowing for localized management, access, and control of the fog network the fog nodes execute data processing[45].

C) Governance: Even when using a private or consortium blockchain, serious issues arise regarding the type of blockchain to deploy (e.g., Hyperledger or Ethereum)[46], who handles and investigates the blockchain, the installation and location of the blockchain nodes, who writes the smart contracts, dispute resolution, the choice of trusted oracles, mechanisms for off-chain activities, side transmit deployment, regulations and standards to comply with, and many others.

The future prospectives of this is an adaptive blockchain design which can be preferred for lightening the computational resource utilization for BCT and 5G network communication for faster services. The cyber physical social system uses blockchain for access control[47].The decentralized and immutable ledger with advanced technologies ensure data integrity and analytics provide better insight for prediction for humongous data accumulation. There are many more potentials which have to be explored. As an example, transaction records are not completely utilized[48], owing to the inability of the application programming interface to be used. Participants with access to transaction big data may have fewer options academic pursuits that are not profit-oriented[49]. A few bitcoin shortcomings, including as energy inefficiency, computational scalability, market entrance obstacles, and regulatory issues, may find adequate remedies through big data analytics. The methods involved with the big data analytics[50] are still optimized framework to secure shared blockchain systems.

6. Conclusion

Blockchain, a distributed ledger platform, has spurred great interest in several big data methods. Its main feature is its efficiency in terms of strong security and a well-functioning network system. This study aimed to cover all aspects of blockchain as a method, as well as its benefits and advantages in large data systems. We attempted to illustrate the rationale for combining these two massive technologies. This study

attempted to emphasize the collecting, storage, analytics and privacy protection of big data. While conducting the research, we identified a number of major challenges that will undoubtedly be discussed further. If a solid blockchain architecture can be built, it will be able to address the problems that come with big data. The large amount of data collected.

Funding

This research received no external funding.

Conflict of interest

The authors declare no conflict of interest.

References

1. Manyika J, Chui M, Brown B, et al. *Big Data: The Next Frontier for Innovation, Competition, And Productivity*. McKinsey Global Institute; 2011.
2. Hwang K, Chen M. *Big-Data Analytics for Cloud, IoT And Cognitive Computing*. Wiley; 2017. pp. 432.
3. Siddiqa A, Hashem IAT, Yaqoob I, et al. A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications* 2016; 71: 151–166. doi: 10.1016/j.jnca.2016.04.008
4. Wu J, Ota K, Dong M, et al. Big data analysis-based security situational awareness for smart grid. *IEEE Transactions on Big Data* 2018; 4(3): 408–417. doi: 10.1109/TBDATA.2016.2616146
5. Tole AA. Big data challenges. *Database Systems Journal* 2013; 4: 31–40.
6. Liu CH, Lin Q, Wen S. Blockchain-enabled data collection and sharing for industrial IoT with deep reinforcement learning. *IEEE Transactions on Industrial Informatics* 2019; 15(6): 3516–3526. doi: 10.1109/TII.2018.2890203
7. Liu G, Dong H, Yan X, et al. B4SDC: A blockchain system for security data collection in MANETs. *IEEE Transactions on Big Data* 2022; 8(3): 739–752. doi: 10.1109/TBDATA.2020.2981438
8. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. In: 2018 Annual National Seminar; 30 May–1 June 2018.
9. Bernabe JB, Canovas JL, Hernandez-Ramos JL, et al. Privacy-preserving solutions for blockchain: Review and challenges. *IEEE Access* 2019; 7: 164908–164940. doi: 10.1109/ACCESS.2019.2950872
10. Parkin J. The senatorial governance of bitcoin: Making (de) centralized money. *Economy and Society* 2019; 48(4): 463–487. doi: doi.org/10.1080/03085147.2019.1678262
11. Casino F, Dasaklis YK, Patsakis C. A systematic literature review of blockchain-based applications: current status, classification and open issues. *Telematics and Informatics* 2019; 36: 55–81. doi: 10.1016/j.tele.2018.11.006
12. Zhang J, Zhong S, Wang T, et al. Blockchain based systems and applications: A survey. *Journal of Internet Technology* 2020; 21: 1–14. doi: 10.3966/160792642020012101001
13. Shalini S, Santhi H. A survey on various attacks in bitcoin and cryptocurrency. In: 2019 International Conference on Communication and Signal Processing (ICCSP); 4–6 April 2019; Chennai, India. pp. 0220–0224.
14. Huang H, Kong W, Zhou C, et al. A survey of state-of-the-art on blockchains: Theories, modelings, and tools. *arXiv* 2020; arXiv:2007.03520. doi: 10.48550/arXiv.2007.03520
15. Dey AK, Akcora CG, Gel YR, Kantarcioglu M. On the role of local blockchain network features in cryptocurrency price formation. *The Canadian Journal of Statistics* 2020; 48(3): 561–581. doi: 10.1002/cjs.11547
16. Angelis J, da Silva ER. Blockchain adoption: A value driver perspective. *Business Horizons* 2019; 62(3): 307–314. doi: 10.1016/j.bushor.2018.12.001
17. Chen Y, Bellavitis C. Blockchain disruption and decentralized finance: The rise of decentralized business models. *Journal of Business Venturing Insights* 2020; 13: e00151. doi: 10.1016/j.jbvi.2019.e00151
18. Moin S, Karim A, Safdar Z, et al. Securing IoTs in distributed blockchain: Analysis, requirements and open issues. *Future Generation Computer Systems* 2019; 100: 325–343. doi: 10.1016/j.future.2019.05.023
19. Zhang R, Xue R, Liu L. Security and privacy on blockchain. *ACM Computing Surveys* 2019; 52(3): 1–34. doi: 10.1145/3316481
20. Leonardos S, Reijsbergen D, Piliouras G. PREStO: A systematic framework for blockchain consensus protocols. *IEEE Transactions on Engineering Management* 2020; 67(4): 1028–1044. doi: 10.1109/TEM.2020.2981286
21. Tang MJ, Alazab M, Luo Y. Big data for cybersecurity: Vulnerability disclosure trends and dependencies. *IEEE Transactions on Big Data* 2019; 5(3): 317–329. doi: 10.1109/TBDATA.2017.2723570
22. Yu H, Yang Z, Sinnott RO. Decentralized big data auditing for smart city environments leveraging blockchain technology. *IEEE Access* 2018; 7: 6288–6296. doi: 10.1109/ACCESS.2018.2888940
23. Rahman MA, Rashid MM, Hossain MS, et al. Blockchain and IoT-based cognitive edge framework for sharing economy services in a smart city. *IEEE Access* 2019; 7: 18611–18621. doi: 10.1109/ACCESS.2019.2896065

24. Dwivedi AD, Srivastava G, Dhar S, Singh R. A decentralized privacy-preserving healthcare blockchain for IoT. *Sensors* 2019; 19(2): 326. doi: 10.3390/s19020326
25. Vyas JD, Han M, Li L, et al. Integrating blockchain technology into healthcare. In: Proceedings of the 2020 ACM Southeast Conference (ACM SE' 20); 2–4 April 2020; New York, NY, USA. pp. 197–203.
26. Jabbari A, Kaminsky P. Blockchain and supply chain management. Available online: https://www.mhi.org/downloads/learning/cicmhe/blockchain-and-supply-chain-management.pdf (accessed on 26 July 2023).
27. Liu CH, Lin Q, Wen S. Blockchain-enabled data collection and sharing for industrial IoT with deep reinforcement learning. *IEEE Transactions on Industrial Informatics* 2019; 15(6): 3516–3526. doi: 10.1109/TII.2018.2890203
28. Xu C, Wang K, Li P, et al. Making big data open in edges: A resource-efficient blockchain-based approach. *IEEE Transactions on Parallel and Distributed Systems* 2019; 30(4): 870–882. doi: 10.1109/TPDS.2018.2871449
29. Sun J, Yao X, Wang S, Wu Y. Blockchain-based secure storage and access scheme for electronic medical records in IPFS. *IEEE Access* 2020; 8: 59389–59401. doi: 10.1109/ACCESS.2020.2982964
30. Li H, Han D. EduRSS: a blockchain-based educational records secure storage and sharing scheme. *IEEE Access* 2019; 7: 179273–179289. doi: 10.1109/ACCESS.2019.2956157
31. Shen M, Zhang J, Zhu L, et al. Secure SVM training over vertically-partitioned datasets using consortium blockchain for vehicular social networks. *IEEE Transactions on Vehicular Technology* 2020; 69(6): 5773–5783. doi: 10.1109/TVT.2019.2957425
32. Andoni M, Robu V, Flynn D, et al. Blockchain technology in the energy sector: A systematic review of challenges and opportunities. *Renewable and Sustainable Energy Reviews* 2019; 100: 143–174. doi: 10.1016/j.rser.2018.10.014
33. Rabah K. Convergence of AI, IoT, big data and blockchain: A review. *The Lake Institute Journal* 2018; 1(1): 1–18.
34. Ionescu L. Big data, blockchain, and artificial intelligence in cloud-based accounting information systems. A*nalysis and Metaphysics* 2019; 18: 44–49. doi: 10.22381/AM1820196
35. Gligor DM, Pillai KG, Golgeci I. Theorizing the dark side of business-to-business relationships in the era of AI, big data, and blockchain. *Journal of Business Research* 2021; 133: 79–88. doi: 10.1016/j.jbusres.2021.04.043
36. Chen J, Lv Z, Song H. Design of personnel big data management system based on blockchain. *Future Generation Computer Systems* 2019; 101: 1122–1129. doi: 10.1016/j.future.2019.07.037
37. Hassani H, Xu H, Silva E. Banking with blockchain-ed big data. *Journal of Management Analytics* 2018. doi: 10.1080/23270012.2018.1528900
38. Baza M, Fouda MM, Nabil M, et al. Blockchain-based distributed key management approach tailored for smart grid. In: Fadlullah Z, Khan Pathan AS (editors). *Combating Security Challenges in the Age of Big Data. Advanced Sciences and Technologies for Security Applications.* Springer, Cham; 2020. pp. 237–263.
39. Deepa N, Pham QV, Nguyen DC, et al. Survey on blockchain for big data: Approaches, opportunities, and future directions. *arXiv* 2021; arXiv:2009.00858. doi: 10.48550/arXiv.2009.00858
40. Lv Z, Liang Q, Hossain MS, Choi BJ. Analysis of using blockchain to protect the privacy of drone big data. *IEEE Network* 2021; 35(1): 44–49. doi: 10.1109/MNET.011.2000154
41. Bhuiyan MZA, Zaman A, Wang T, et al. Blockchain and big data to transform the healthcare. In: Proceedings of the International Conference on Data Processing and Applications (ICDPA 2018); 12–14 May 2018; Guangzhou China. pp. 62–68.
42. Xu C, Wang K, Li P, et al. Making big data open in edges: A resource-efficient blockchain-based approach. *IEEE Transactions on Parallel and Distributed Systems* 2019; 30(4): 870–882. doi: 10.1109/TPDS.2018.2871449
43. Bandara E, Ng WK, De Zoysa K, et al. Mystiko—Blockchain meets big data. In: 2018 IEEE International Conference on Big Data (Big Data); 10–13 December 2018; Seattle, WA, USA. pp. 3024–3032.
44. Lv Z, Qiao L, Hossain MS, Choi BJ. Analysis of using blockchain to protect the privacy of drone big data. *IEEE Network* 2021; 35(1): 44–49. doi: 10.1109/MNET.011.2000154
45. Xu C, Wang K, Li P, et al. Making big data open in edges: A resource-efficient blockchain-based approach. *IEEE Transactions on Parallel and Distributed Systems* 2018; 30(4): 870–882. doi: 10.1109/TPDS.2018.2871449
46. Li J, Wu J, Jiang G, Srikanthan T. Blockchain-based public auditing for big data in cloud storage. *Information Processing & Management* 2020; 57(6): 102382. doi: 10.1016/j.ipm.2020.102382
47. Salah K, Ur Rehman MH, Nizamuddin N, Al-Fuqaha A. Blockchain for AI: Review and open research challenges. *IEEE Access* 2019; 7: 10127–10149. doi: 10.1109/ACCESS.2018.2890507
48. Thangaraj M, Suguna S, Sudha G. *Big Data Analytics: Concepts, Techniques, Tools and Technologies*. PHI Learning Private Limited; 2022. pp. 416
49. Guo L, Xie H, Li Y. Data encryption based blockchain and privacy preserving mechanisms towards big data. *Journal of Visual Communication and Image Representation* 2020; 70: 102741. doi: 10.1016/j.jvcir.2019.102741
50. Marichamy VS, Natarajan V. Blockchain based securing medical records in big data analytics. *Data & Knowledge Engineering* 2023; 144: 102122. doi: 10.1016/j.datak.2022.102122